DNA2FACE: An Approach to Correlating 3D Facial Structure and DNA *

Nisha Srinivas, Ryan Tokola, Aravind Mikkilineni, Intawat Nookaew,
Michael Leuze, and Christopher Boehnen
Oak Ridge National Laboratory
{srinivasn, tokolara, mikkilinenak, nookaewi, leuzemr, boehnencb}@ornl.gov

Abstract

In this paper we introduce the concept of correlating genetic variations in an individual's specific genetic code (DNA) and facial morphology. This is the first step in the research effort to estimate facial appearance from DNA samples, which is gaining momentum within intelligence, law enforcement and national security communities. The dataset for the study consisting of genetic data and 3D facial scans (phenotype) data was obtained through the FaceBase Consortium. The proposed approach has three main steps: phenotype feature extraction from 3D face images, genotype feature extraction from a DNA sample, and genome-wide association analysis to determine genetic variations that contribute to facial structure and appearance. Results indicate that there exists significant correlations between genetic information and facial structure. We have identified 30 single nucleotide polymorphisms (SNPs), i.e. genetic variations, that significantly contribute to facial structure and appearance. We conclude with a preliminary attempt at facial reconstruction from the genetic data and emphasize on the complexity of the problem and the challenges encountered.

1. Introduction

The reduction in cost of DNA sequencing and genotyping technology is leading to the availability of massive datasets of human genome sequences and human genotypes representing genetic variations among people. This provides for the first time the potential to correlate an individual's specific genetic code with susceptibility to disease, behavior, physical appearance, and other expressed traits. Facial morphology is affected by a combination of genetic, environmental, and possible epigenetic factors. However, an individual's DNA is arguably the most significant factor in determining facial structure and appearance, which is easily evidenced by comparing children to their parents or examining identical twins.

Currently, DNA based recognition is the universally accepted approach for determining the identity of an individual. However, there does not exist a single database composed of everyone's identity and DNA information. In many criminal cases there exists DNA evidence but no database sample for comparison to enable identification. In such situations it is extremely effective to have a technology that could predict facial structure and appearance from the DNA samples to search for a suspect via methods such as wanted posters or potentially using biometrics to search facial recognition databases. The idea of predicting facial structure and appearance from DNA samples has caught the eye of many researchers [14, 18, 17, 12], however the science behind it is still in it's early stages and has not yet shown itself to be of high enough quality for use as credible evidence.

The primary objective of this work is to introduce the biometrics community to the current state of genetic facial prediction, and to show that there exists significant correlations between genetic information and facial structure. The existence of these correlations is important given that the majority of a person's genetic code is likely unrelated to facial structure. This also suggests that it should be possible to develop appropriate machine learning methods to predict either facial structure from genetic information, or vice versa. We conclude with a preliminary attempt at facial reconstruction from the genetic data and discussion of the results. Specifically we will emphasize the complexity of the problem and provide insight into the current challenges and future research that will need to be performed to improve the state of the art.

The paper is organized as follows: Prior research in cor-

*This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy.

relating facial appearance with genetic data is first summarized in Section 2. Details of the proposed approach and how it differs from prior and existing methods are described in Section 3. Section 4 describes the genetic and 3D face dataset used for this study. Finally the Genome Wide Association Study (GWAS) that is performed on the dataset and its associated results are described in Section 5. In Section 5.1 we then present an initial study on predicting facial structure from the identified genes that affect facial morphology. The paper concludes with a brief summary and discussion.

2. Related Work

Recent studies on facial development have started to identify genes involved in defining facial morphology, many of which work together in a concerted effort [15]. The lack of acknowledgment of these inter-dependencies is a major failing of traditional methods for modeling facial morphology, which have focused on individual characteristics viewed in isolation, such as nose length, chin size, or cheekbone width [18]. While these are simple, easily defined, and measurable characteristics, the traits themselves are almost certainly related to each other and to multiple shared underlying factors. Part of the reason why the interdependent nature of these features has not been closely studied is the lack of statistical and machine learning methods that are feasible given the large size of the genomic data sets involved. Some minimal inter-dependency studies have been performed to predict a person's ethnic origin by combining faces of the corresponding ethnicities [17] identified through existing genetic analysis methods. While such ethnic information is useful, it is still limited to identifying a large population group and not a specific individual.

3. Proposed Approach

As stated in the introduction, the objective of this work is to identify genes associated with facial morphology. Our proposed approach to this problem is shown in figure 1 and explained in detail in Sections 3.1, 3.2, and 5. We begin with two inputs: 3D scans of faces (phenotype), and associated biological samples from the subjects which are sequenced to produce a set of genetic data (genotype). The 3D scans are aligned with a *reference face*, and a comparison between the scanned face and the reference face produces a *correspondence vector* which is a mathematically-defined representation of the facial structure. Principal component analysis (PCA) [16] is applied to the correspondence vectors resulting in a small set of *scores*. The PCA scores are then compared to the genetic data using a method known as Genome Wide Association Studies (GWAS) [23, 9]. The result is a statistically significant set of genetic variations that are associated with facial morphology. Throughout this work the terms genotype and phenotype will be used. Genotype refers to the set of genes that an individual carries, while phenotype refers to the observable characteristics expressed by those genes.

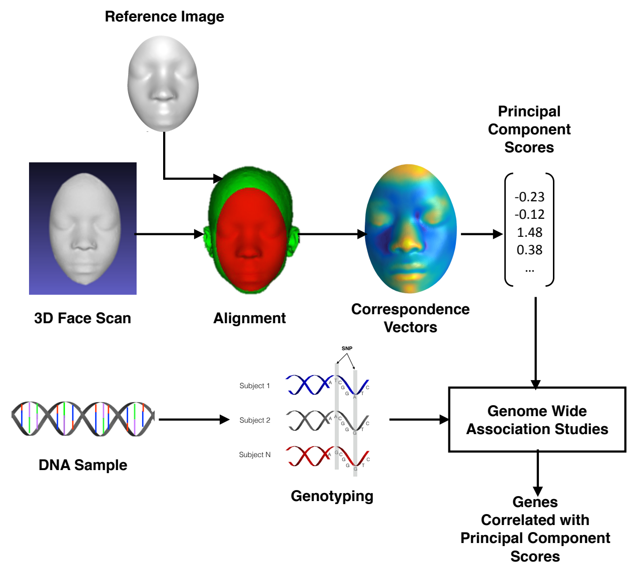

Figure 1: Proposed approach to finding correlations between DNA and facial structure.

3.1. Extracting 3D Facial Features: Phenotype Data

In previous work the phenotype data that was used to represent facial morphology focused on individual features such as nose length or eye separation. In our proposed approach we define a feature vector that captures variation across the entire face simultaneously. We extract 3D signatures known as correspondence vectors (CV) [7] from each subject's face. A CV measures the distance between an input face and a reference face, mapping points from a global coordinate system to face-based coordinates. The CV transforms a 3D vector (x, y, z) into a 1D vector (d), resulting in decreased computational complexity and a simpler PCA calculation.

The main steps in calculating correspondence vectors are:

3.1.1 Reference Face Generation

A reference face specific to the Tanzanian dataset is created using the procedure outlined in [7]. It is used to establish a canonical face-coordinate system. Initially, N 3D face scans are randomly sampled from the dataset, where $N = 94$. An initial reference consisting of 15,000 points was generated by aligning, averaging, smoothing, and uniformly sampling ten of the N 3D face scans. We used the Raindrop Geomagic software package [3] for this task. The remaining 84 images are used as a training set to compute the final reference image. Generating the final reference image is an iterative process. In each iteration, for each image in the training set the initial reference image is aligned with the image and the correspondence vector is calculated using the distance along the surface normal. Next, the correspondence vectors are averaged across the training set to create an average correspondence vector. This is then mapped back to a surface and is used as the initial reference face for the next iteration. The process is then repeated until the convergence criteria was met.

3.1.2 Alignment

We align the reference face and the input face in a two stage process:

1. **Initial alignment:** The input face is first translated and rotated such that it "faces" the $-y$ direction and has the tip of the nose located at the origin $(0,0,0)$.

2. **Fine alignment:** The initial alignment is used as an initialization for the iterative closest point algorithm (ICP) [6]. It is a gradient-based iterative method that minimizes the overall distance between the two face surfaces.

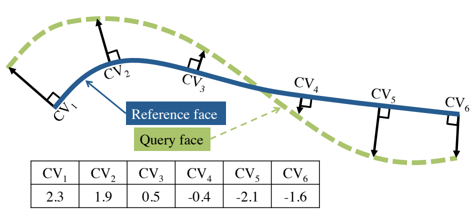

Figure 2: Schematic illustration of a CV with six elements.

3.1.3 Calculating CVs

The reference face and input faces are surfaces defined by a set of spatial coordinates (vertices) and a connectivity map describing how those points are connected to form the surface. It is important to note that there is no common sampling grid between any two faces that would allow a simple difference to be taken between the two surfaces to obtain a signature. Instead a "correspondence" is found between the two surfaces that is analogous to a pointwise distance given a common sampling grid. Figure 2 illustrates the correspondence problem. The blue line represents the reference surface, and green dotted line represents the input surface. The vertices of the reference surface are taken to be the sampling grid and a normal search approach is used to determine the correspondence of the input surface at that point. The set of correspondences is called the correspondence vector. The normal search method defines the correspondence at a specific vertex $\mathbf{v}_r$ of the reference surface to be the point of intersection between the input surface and the reference surface normal at $\mathbf{v}_r$. The set of correspondence points can be ordered based on a predefined ordering of the reference vertices and the resulting ordered set can be thought of as a vector of correspondence points or a correspondence vector. Each correspondence point can be replaced with its distance from the corresponding reference vertex to reduce the correspondence vector to 1-dimension without loss of information. This is possible since the correspondence points for all input images will always lie along the same set of normal vectors defined by the reference surface.

Thresholding of CV : Even with a good alignment, a CV will occasionally have a large "outlier" value. This typically happens when there is a hole in the input face mesh, and is most common around nostrils. Very large values in CV can significantly impact statistical studies, such as GWAS. Outliers are removed by applying a hard threshold to CV. The threshold prevents large values from skewing the statistical analysis and is completely automated.

PCA of CV: To more efficiently compare faces, it is common to use statistical methods such as PCA to reduce the dimensionality of the data. Approaches such as PCA create a set of basis vectors. The idea is that any face can be represented as a weighted linear combination of basis vectors. In the specific case of PCA, these basis vectors are also called principal components (PC). The linearity is important, as it allows a face to be reconstructed given the basis vectors and a set of weights. It is these weights that represent facial appearance (phenotype) and is used to find genes correlated to facial morphology.

Applying PCA to the Tanzanian dataset results in a set of PCs that together describe the full set of CVs. Since the Tanzanian dataset contains 3555 images which is less than the CV dimensionality of 7371, the number of PCs obtained is 3555. Every CV can be perfectly reconstructed using a weighted sum of the PC: $\psi_j = \sum_i \alpha_{j,i}\phi_i = \alpha_j\Phi$. For example, the CV of the 17th image in the dataset can be represented as $\psi_{17} = 0.53\phi_1 - 0.03\phi_2 + 0.70\phi_3 + \cdots + 0.19\phi_{3556}$ where ϕ_i represents the i^{th} PC. In general each CV can then be fully represented by a vector α such that $\psi_j = \alpha_j\Phi$.

The PCs not only provide a way to represent each CV in the training set, they can be used to represent CVs from any face. Once the PCs have been created, any face can be represented by a vector α, and a 3D model of the face can be reconstructed from the resulting CV. The accuracy of the reconstruction depends only upon how well the face is represented in the training data. We only consider the first 30 PCs as it captures 90% of the variance in the dataset. Figure 3 shows a graphical representation of the first twenty PCs from the Tanzanian dataset. The elements of each PC are color mapped with yellow indicating more positive values and blue indicating more negative values. Since the original CV can be interpreted as a set of distances from a standard reference face, the PCs have an an intuitive meaning in the sense that each element indicates whether that point lies in front of or behind the reference face. Regions where the PC is positive (yellow) valued will tend to be in front of the reference face whereas regions where the PC is negative (blue) tend to be behind the reference face. Consider the third PC shown in Figure 3. A CV that results from multiplying this PC with a positive value means that the nose protrudes more than with an average face, and the cheeks protrude less. Similarly a CV that results from multiplying the twelfth component by a large positive value makes a face with a broader nose.

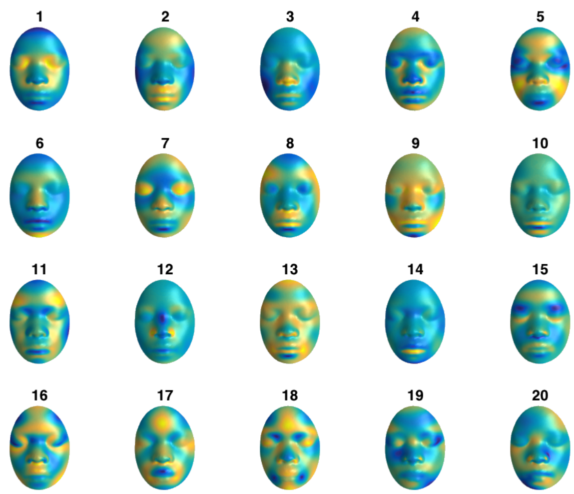

Figure 3: Visualization of first 20 PC from the dataset.

3.2. Extracting Genetic Features

Humans are made up of trillions of cells. The nucleus in each cell consists of 23 pairs of chromosomes; 22 autosomal (non-sex chromosomes) pairs and the sex chromosome pair, XX for female and XY for male. In each pair, one chromosome is inherited from the mother and other from the father; hence humans are diploid organisms. Chromosomes are made of a deoxyribonucleic acid (DNA) molecules and proteins. DNA is a double helix structure and each helix is made up of a series of nucleotides as illustrated in Figure 4. Each nucleotide is a complex molecule comprised of a base, sugar, and phosphate group. When discussing the DNA sequence we typically consider only the base of each nucleotide. The four bases are adenine (A), guanine (G), cytosine (C) and thymine (T). The bases always appear in pairs between both strands of the helix; A with T and C with G [8]. It is the order or sequence of these nucleotides along a single helix that form the genetic code which is referred to as the genome. Short sections of the DNA are known as genes which are the fundamental units of hereditary. Each gene is present at specific locations (loci) on the chromosomes across the human population and determine the characteristics of an individual. The Human Genome Project [21] has identified approximately $25,000$ genes in humans. Different versions of the same gene are what makes a person unique. Since the chromosomes in human cells appear in pairs, there exist two copies of the same gene, one inherited from each parent and located at the same location on each chromosome pair. However, the two gene copies may differ.

Although the human genome consists of approximately 6.5 billion base pairs, only a relatively small subset of these ($<1\%$, or approximately 10 million) exhibit variation between individual humans. These points in the genome where variations occur among the population are referred to as Single Nucleotide Polymorphisms (SNPs) [24, 1], and the different variants of the SNP are called alleles. SNPs are typically identified by their position within the genome, including which specific gene, if any, they are part of. A process called genotyping is used to measure SNPs from a DNA sample such as saliva or blood [13].

Figure 5 illustrates SNPs as they would differ among a population. In the figure, variations at the nucleotide level in the DNA sequence between three subjects are observed at two different locations. SNP_1 is said to have two alleles A and T, implying that within a population there are two commonly occurring bases (A and T) at that particular location, with A occurring in majority of the population (major allele) and T appearing in the remaining population (minor allele). Similarly, SNP_2 is said to have two alleles G (major allele) and A (minor allele).

4. Dataset

In this study, we use features extracted from 3D face scans to represent the phenotype data and the genotype data is expressed in the form of variations of SNPs at known locations. The dataset was obtained from the FaceBase repository [2] at the National Institute of Dental and Craniofacial Research (NIDCR) [4]. The dataset was collected over a three year period in the Mwanza region of Tanzania under the NIDCR FaceBase initiative. The dataset consists of $3,623$ unique subjects from Tanzania. Only $3,557$ subjects have both genotype and phenotype data available. Figure 6 shows the age distribution of subjects in the dataset. Approximately 55.5% of subjects were female and 44.5% were male. The age range of the subjects varied between 3 to 21 and followed an approximately normal distribution with an average age of 10.7. The ethnicity of the subjects is presented in Table 1 . We only consider subjects of Bantu-Bantu ethnicity in this study as a homogeneous (mono-ethnic) dataset is more likely to show consis-

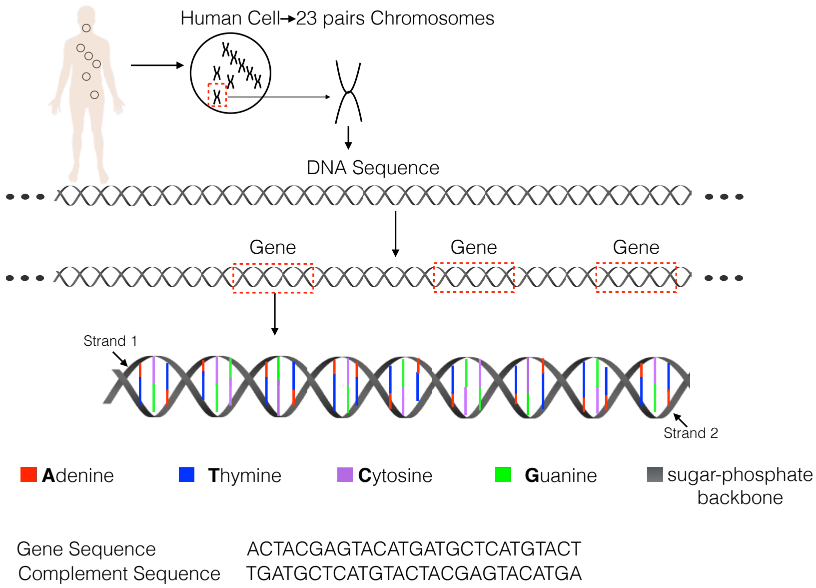

Figure 4: Overview of the relationship between DNA and genes.

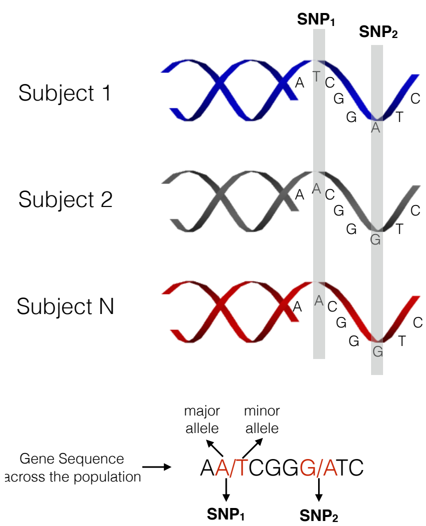

Figure 5: Single nucleotide polymorphisms at different locations (locus) on the gene across a sample population.

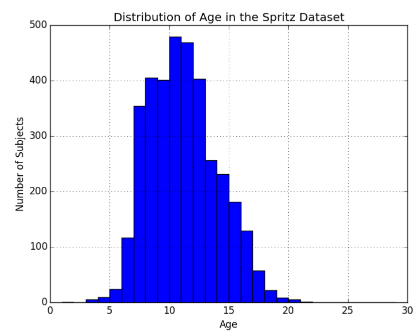

Figure 6: Age distribution of subjects in the dataset.

Mom-Dad Ethnicity	Number of Subjects
Bantu-Bantu	3524
Bantu-Nilotes	15
Nilotes-Bantu	1
Bantu-Cushitic	1
Cushitic-Cushitic	1
Bantu-Hamites	1
No_Data-Bantu	1
Bantu-No_Data	1
No_Data-No_Data	10

Table 1: Ethnicity distribution in the Dataset.

tent genotype/phenotype associations.

4.1. Phenotype Data: 3D Face Scans

The dataset consists of $3,623$ 3D face scans captured using the Creaform MegaCapturor (MC) camera three-dimensional (3D) photogrammetry imaging system. Each face is oriented such that it is facing in the $-y$ direction with the origin of the coordinate system located in front of the facial plane. Additional phenotype data provided for each subject include age, sex, height, weight and ethnicity information.

Subject ID	SNP_1	SNP_2	SNP_3	..	SNP_N
Subject 1	QQ	Qq	QQ	..	Qq
Subject 2	Qq	qq	Qq	..	Qq
Subject 3	qq	QQ	qQ	..	Qq
Subject 4	qQ	qQ	QQ	..	Qq
:	:	:	:	..	:
Subject M	qq	Qq	QQ	..	Qq

Table 2: Scheme used for representing measured SNPS.

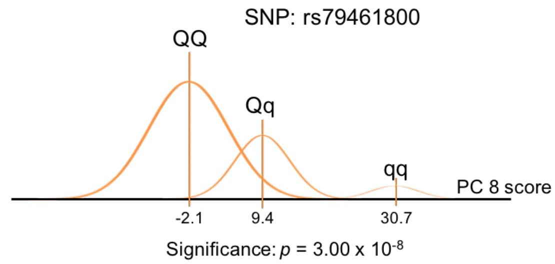

Figure 7: Quantitative trait analysis using GWAS

4.2. Genotype Data

The available genotype data for each individual in the FaceBase dataset is expressed in the form of SNPs obtained by genotyping a DNA sample obtained from saliva samples. FaceBase genotyping uses Illumina technology, specifically a HumanOmni25Exome array, where 25 indicates that measurements are attempted at 2.5 million locations in the genome. However, only $701,928$ SNPs were recorded with non-missing values in the dataset.

Measured SNPs for subjects in a dataset are expressed in the format shown in table 2. Since humans are diploid organisms and have two copies of the same genes located at the same location on each chromosome pair, SNP values are expressed in terms of allele pairs; QQ, Qq or qq; QQ indicates that the major allele is present at a given SNP in both chromosomes in the pair, Qq indicates that the major allele is present in one chromosome and the minor allele in the other, and qq indicates the presence of the minor allele in both chromosomes.

5. GWAS Analysis and Results

A Genome Wide Association Study (GWAS) identifies associations between a set of commonly occurring genetic variants and some observable trait. The methods used to identify these associations typically examine the variance in the phenotype distribution for a given SNP. GWAS has been most commonly applied to human disease, comparing case and control groups. In our case we use GWAS to find genomic associations with facial morphology, or more specifically between a set of SNPs and the weightings of the principal components (PCs) described in Section 3.1.

Consider the example of finding an association between SNP_1 and the first principal component (PC_1). To perform the association testing the mean and variance are obtained for the phenotype data associated within each allele group, Qq, QQ, and qq, of the SNP. This is shown in figure 7. Next, statistical tests are used to obtain the associated p-values for the distribution of the phenotype data across allele group for the given SNP. A high p-value implies low association between the SNP and the phenotype meaning that the SNP variants do not influence the phenotype expression. A low p-value implies a high association between the SNP and the phenotype expression. An association with a p-value of less than $5 \times 10^{-}8$ is known as genome-wide significant by the GWAS community and a p-value of less than $1.0 \times 10^{-}5$ is known as "suggestive of significance" [19]. We used PLINK [20], an open source tool to perform the genome wide association study.

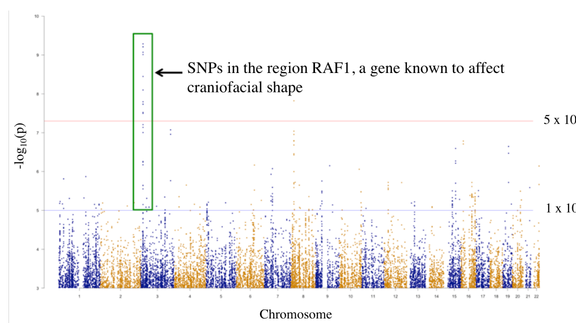

Figure 8: Manhattan plot of SNP significance in a GWAS of African subjects.

GWAS associations were found between each principal component and the 701928 SNPs recorded during genotyping. The analysis identified 30 SNPs with p-values below 5.0×10^{-8} and 5628 SNPs with p-values below 1.0×10^{-5} across all 30 principal components. This indicates that there exists significant correlations between facial features and genetic variations.

Most of the identified SNPs occur in clusters that are associated with individual genes or small groups of contiguous genes. The most significant cluster of variations, with p-values smaller than 1.0×10^{-9}, is found on chromosome 3 in the region of RAF1, a gene known to affect craniofacial shape, as shown in Figure 8. In the first 30 principal components, there are approximately 70 other clusters with significance that extends well into the "suggestive of significance" range, with p-values smaller than 5.0×10^{-7}. Most of these clusters are in the regions of genes not previously known to be associated with facial morphology. Many of these genes produce protein kinases or protein phosphates, which are enzymes that add or remove phosphate groups from other proteins and thereby modify their function. They are the enzymes that regulate the majority of cellular pathways. We hypothesize that since the "energy" of a PC is typically spread across the face rather than localized to a particular region we are discovering genes that regulate other genes which affect facial shape. This is an encouraging result suggesting an approach to discovering components and interactions of the gene regulatory networks (GRNs) that affect facial shape.

5.1. Predicting Facial Structure from DNA: Preliminary Results

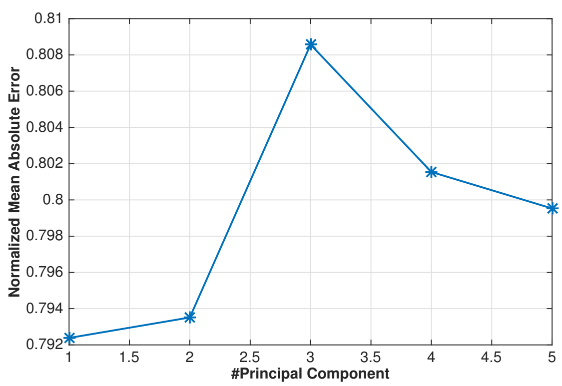

Figure 9: Mean 5-fold cross validation results for the top 5 principal components.

The primary goal of this paper was to demonstrate the existence of genetic/facial correlations which was shown in the previous section. Identifying appropriate correlates is critical to being able to extend these results to prediction of facial features (principal component scores) solely from genomic information.

A preliminary attempt at facial prediction is performed using the set of SNPs identified in the previous section. The problem is formulated as a regression problem to predict the principal component scores given a set of SNP variants.

We use support vector regressors (SVRs) [5] to predict the principal component scores (phenotype data) from the most significant SNPs (genotype data). The LIBSVM software package [10] is used with a radial basis function (Gaussian) kernel for all our SVR experiments. As with SVMs, the optimal C and γ parameters were found with a grid search in a nested 5-fold cross-validation framework. In order to allow comparisons between principal components, results are shown in terms of normalized absolute error where the average absolute error is divided by the standard deviation of the principal component values in the dataset. Results for the first 5 principal components are shown in Figure 9. The scatter plots of actual vs. predicted values for each cross validation *PC* is shown in figure 10. From the results we infer that the trained models tended to predict the average value for each principal component.

Although the regressor in this case was unable to generate a strong predictive model, we have identified several limiting factors that can be addressed in future studies:

1. Age variation: The age variability in the dataset is a significant source of noise. Older children tend to have larger faces than younger children and this size difference can cause the PCs to spend significant energy capturing size over other spatial features. Using a more homogeneous dataset can remove some of this undesired variability in the PCs.

2. Limited data: Genomic problem in general are extremely complex, however the addition of high-dimensionality phenotype data adds an additional layer of complexity when performing GWAS compared to more typical binary disease/non-disease type studies. Machine learning algorithms perform best when the number of data samples greatly exceeds the dimensionality of the data. Non-traditional GWAS methods may need to be developed to deal with this problem in an appropriate manner.

3. Genomic complexity: The complexity of the human genome is well understood. A system that completely predicts facial structure will most likely need to incorporate a large amount of genetic data beyond just a small number of SNPs. This kind of analysis will require significant computational resources and methods for reducing the computational complexity of the algorithms involved.

4. Regulatory genes: One of the disadvantages of traditional GWAS is that only 1-to-1 relationships between genotype and phenotype are considered. Any potential correlation between SNPs themselves are ignored. Genes, however, do not act in isolation. Some genes, such as *regulatory* genes, are not directly expressed, but instead modify the level of expression of other genes. Traditional GWAS methods do not search for or find these relationships. One way to find appropriate correlations with both kinds of SNPs is to modify GWAS to operate on multiple SNPs simultaneously.

5. Incomplete and categorical data: One way in which genetic data differs from the input to most traditional machine learning algorithms is that it is essentially categorical: each SNP will either be QQ, Qq, or qq. In addition, not every SNP is genotyped for every subject, so there is also a "missing" category. Many machine learning algorithms, such as SVMs and SVRs, work best when the input has continuous values and is never

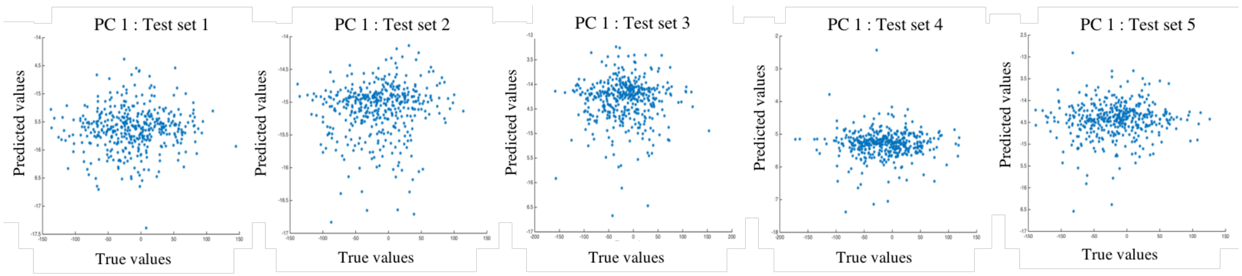

Figure 10: Scatter plot of actual vs predicted value for principal component 1.

missing, but other algorithms, such as decision-based learners, can work well with categorical data.

6. Conclusion

In this work we have shown the existence of genetic markers, SNPs, that exhibit high correlation with facial structure. Unlike previous studies that have attempted to find correlations between SNPs and individual facial features such as nose length or eye separation, we have developed a spatial feature set using PCA to encode full-face variations under the assumption that the mapping from genetic variations to individual facial structures is not a 1-to-1 mapping as evidenced in [15]. We believe further evidence of this is given by the GWAS results obtained in this paper which identify a large number of regulatory genes as having significant correlation with facial structure. By their nature, regulatory genes affect many other genes that together affect the expressed facial structure.

The preliminary attempt to reconstruct a face given the genetic data, although not successful at this stage, provided additional insight into the underlying challenges as described in Section 5.1. Several of these challenges can be addressed to some degree by data pruning together with operation on a significantly larger data set. The majority of the challenges however have to do with the available statistical and machine learning methods and their limitations in terms of dealing with high dimensionality data, mix of categorical and continuous features, and most importantly computational feasibility. Recent work in areas such as genomic compression [22, 11] may provide insights into appropriate algorithm development to address the computational challenges.

References

[1] 23andme. https://www.23andme.com.

[2] Facebase: A resource for craniofacial researchers. `https://www.facebase.org/`.

[3] Geomagic studio. `http://www.geomagic.com/en/products-landing-pages/3d-design/`.

[4] National institute of dental and craniofacial research. `https://www.nidcr.nih.gov`.

[5] S. Abe. *Support vector machines for pattern classification*, volume 2. Springer, 2005.

[6] P. J. Besl and N. D. McKay. Method for registration of 3-d shapes. In *Robotics-DL tentative*, pages 586–606. International Society for Optics and Photonics, 1992.

[7] C. Boehnen, T. Peters, and P. Flynn. 3D signatures for fast 3D face recognition. In M. Tistarelli and M. Nixon, editors, *Advances in Biometrics*, volume 5558 of *Lecture Notes in Computer Science*, pages 12–21. Springer Berlin Heidelberg, 2009.

[8] T. A. Brown. *Genomes*. Garland science, 2006.

[9] W. S. Bush and J. H. Moore. Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 2012.

[10] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[11] B. G. Chern, I. Ochoa, A. Manolakos, A. No, K. Venkat, and T. Weissman. Reference based genome compression. In *2012 IEEE Information Theory Workshop*, pages 427–431, Sept 2012.

[12] P. Claes, D. K. Liberton, K. Daniels, K. M. Rosana, E. E. Quillen, L. N. Pearson, B. McEvoy, M. Bauchet, A. A. Zaidi, W. Yao, et al. Modeling 3d facial shape from dna. *PLoS Genet*, 10(3):e1004224, 2014.

[13] . G. P. Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[14] J. Fagertun, K. Wolffhechel, T. H. Pers, H. B. Nielsen, D. Gudbjartsson, H. Stefansson, K. Stefansson, R. R. Paulsen, and H. Jarmer. Predicting facial characteristics from complex polygenic variations. *Forensic Science International: Genetics*, 19:263–268, 2015.

[15] L. G. Farkas, M. J. Katic, and C. R. Forrest. International anthropometric study of facial morphology in various ethnic groups/races. *Journal of Craniofacial Surgery*, 16(4):615–646, 2005.

[16] I. Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[17] Y. C. Klimentidis and M. D. Shriver. Estimating genetic ancestry proportions from faces. *PLoS One*, 4(2):e4460, 2009.

[18] F. Liu, F. Van Der Lijn, C. Schurmann, G. Zhu, M. M. Chakravarty, P. G. Hysi, A. Wollstein, O. Lao, M. De Bruijne, M. A. Ikram, et al. A genome-wide association study

identifies five loci influencing facial morphology in europeans. *PLoS Genet*, 8(9):e1002932, 2012.

[19] O. A. Panagiotou, J. P. Ioannidis, et al. What should the genome-wide significance threshold be? empirical replication of borderline genetic associations. *International journal of epidemiology*, 41(1):273–286, 2012.

[20] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[21] M. P. Sawicki, G. Samara, M. Hurwitz, and E. Passaro. Human genome project. *The American journal of surgery*, 165(2):258–264, 1993.

[22] K. Tatwawadi, M. Hernaez, I. Ochoa, and T. Weissman. Gtrac: fast retrieval from compressed collections of genomic variants. *Bioinformatics*, 32(17):i479, 2016.

[23] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[24] M. P. Weiner and T. J. Hudson. Introduction to snps: discovery of markers for disease. *Biotechniques*, 32:s5–s32, 2002.